SUB-LEXICAL LANGUAGE MODELS FOR GERMAN LVCSR

Amr El-Desoky Mousa, M. Ali Basha Shaik, Ralf Schlüter, Hermann Ney

Lehrstuhl für Informatik 6 – Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany

ABSTRACT

One of the major difficulties related to German LVCSR is the rich morphology nature of German, leading to high out-of-vocabulary (OOV) rates, and high language model (LM) perplexities. Normally, compound words make up an essential fraction of the German vocabulary. Most compound OOVs are composed of frequent in-vocabulary words. Here, we investigate the use of sub-lexical LMs based on different approaches for word decomposition, namely supervised and unsupervised decomposition, as well as decomposition derived from grapheme-to-phoneme (G2P) conversion. In the later approach, we augment a normal word model with a set of grapheme-phoneme pairs called *graphones* used to model the OOV words. A novel approach is proposed to select the representative graphone sequences for OOVs based on unsupervised decomposition and word-pronunciation alignment. We obtain relative reductions in word error rate (WER) from 4.2% to 6.5% with respect to a comparable full-words system.

Index Terms— Speech recognition, language model, sub-lexical, graphone, German

1. INTRODUCTION

German is characterized by high lexical variety as a large number of distinct lexical forms can be generated due to different factors like word compounding, inflection, and derivation. This high lexical variety causes data sparsity problems, and leads to high OOV rates, and high LM perplexities. The traditional way to overcome this problem is to use a large recognition lexicon typically having several hundred thousands of full-words. However, relatively high OOV rates are still measured. In addition, the ASR system suffers from high resource requirements such as CPU time and memory.

For the above reasons, sub-lexical LMs based on word decomposition into fragments are used in order to lower the OOV rate and perplexity, reduce data sparsity, decrease the resource requirements, and improve the final WER as well. Broadly speaking, There are two main approaches to word decomposition, namely, supervised and unsupervised approaches. Normally, the supervised approaches are based on linguistic knowledge like in [1], where a set of about 340 decomposition rules has been manually developed for splitting compound German words. While, in [2], a hand corrected lexicon is used for recognition, where compound words are manually decomposed. Other supervised methods rely on morphological analysis based on lexical and syntactic knowledge like in [3, 4, 5]. Although the supervised splitters are normally optimized for high performance, they require labor-intensive work and still suffer from the so-called *unknown word problem*, that is, words that are not coded into the system. Moreover, a morphology-based splitter may not be readily available. On the other hand, the unsupervised approaches are data-driven statistical-based approaches like in [6], where a set of 800k decomposition rules are automatically extracted. While in [7], a compound splitting algorithm is developed that uses sorting, word length, and word frequency information. In [8], a compound word splitting algorithm is proposed that splits compounds according to the statistical relevance of the resulting constituents. Other unsupervised methods are based on the minimum description length principle (MDL) like in [9]. On the contrary, the unsupervised approaches do not require any language specific knowledge and can be applied to any language.

Another approach used to cope with the high OOV rates, is to augment the traditional word model with a specialized model for OOVs. The goal of this OOV modeling is to be able to spell new words. Moreover, the presence of the OOV words can affect the adjacent words leading to more misrecognition. According to [10], each OOV causes 1.5 to 2 errors on average. The augmenting OOV model can be a phone-based model, so that any OOV word can be recognized as an arbitrary sequence of phonemes like in [9, 11]. In [12], multi-phoneme fragments are automatically constructed and integrated into the lexicon and the language model, but no attempt is made to convert phoneme sequences into proper words. Alternatively, in [13], a set of automatically derived fragments joint with their pronunciations is augmented to the word model leading to small improvements in WER. While, in [10] a graphone-based model is used, where every OOV word is represented as a joint sequence of constrained-size grapheme-phoneme pairs called *graphones*, which are automatically derived from (G2P) conversion [14].

In this work, we investigate the use of sub-lexical LMs for German large vocabulary and continuous speech recognition (LVCSR). We try both supervised and unsupervised word decomposition. On the other hand, we examine LMs built on se-

quences of in-vocabulary words with interspersed sequences of graphones replacing the OOV words. We examine both constrained-size as well as free-size graphones.

The paper is organized as follows: In Section 2, we present the details of our methods. In Section 3, we describe our experimental setup. Experiments are discussed in section 4, while Section 5 draws conclusions.

2. METHODOLOGY

2.1. Pronunciation generation

For words and fragments whose pronunciations are unknown, we use a statistical G2P approach to get the missing pronunciations. Our approach is based on joint-sequence models as shown in [14]. Therein, we search for the most likely pronunciation $\varphi \in \Phi^*$ for a given orthographic form $g \in G^*$, where Φ and G are the sets of phonemes and letters respectively:

$$\varphi(g) = \arg\max_{\acute{\varphi} \in \Phi^*} p(\acute{\varphi}, g) \quad (1)$$

We refer to the joint probability distribution $p(\varphi, g)$ as a "graphonemic" joint sequence model. We assume that for each word, its orthographic form and its pronunciation are generated by a common sequence of graphonemic units called *graphones*. Each graphone is a pair $q = (g, \varphi) \in Q \subseteq G^* \times \Phi^*$ of a letter sequence and a phoneme sequence of possibly different lengths. The joint probability distribution $p(\varphi, g)$ is reduced to a probability distribution over graphone sequences $p(q)$ which are modeled by a standard M-gram:

$$p(q_1^N) = \prod_{i=1}^{N+1} p(q_i | q_{i-1}, ..., q_{i-M+1}) \quad (2)$$

This model has two parameters: the order of the M-gram model, and the allowed size of graphones. The number of letters and phonemes are allowed to vary between zero and an upper limit L. Such a model can be trained using maximum likelihood (ML) training via expectation maximization (EM) algorithm as presented in [14]. To produce a pronunciation for a given word, we use the maximum approximation over the set $S(g, \varphi)$ of all joint segmentations of g and φ:

$$p(\varphi, g) \approx \max_{q \in S(g,\varphi)} p(q_1, ..., q_L) \quad (3)$$

2.2. Word decomposition approaches

2.2.1. Supervised word decomposition

Here we perform German word decomposition based on a predefined set of around 100 common German prefixes and suffixes, along with a set of additional unsupervised splittings for around 17k compound words automatically performed using frequency-based splitting proposed in [15]. It should be noted that this is not a complete supervised approach since some unsupervised decomposition takes place. The frequency-based splitting is done in the following way:

- Each word which consists of two or more in-vocabulary words is considered as a compound word;
- For each compound word w:
  - compute frequency $N(w)$ and component frequencies $N(w_1), ..., N(w_K)$.
  - compute geometric mean of component frequencies $GM(w_1, ..., w_K) = [\prod_{k=1}^{K} N(w_k)]^{\frac{1}{K}}$
  - split word w if $GM(w_1, ..., w_K) > N(w)$

All the LM training corpora are processed such that the predefined prefixes and suffixes are stripped off, and the compound words are decomposed.

2.2.2. Unsupervised word decomposition

Here we perform German word decomposition based on unsupervised techniques implemented by *Morfessor* [16, 17]. It is a tool that works in an unsupervised manner, and autonomously discovers segmentations for the words in unannotated text corpora. Moreover, it is a general model for unsupervised induction of morphology from raw text. It is designed to cope with languages having predominantly a concatenative morphology, and where the number of morphemes per word is varying so much and not known in advance. Although Morfessor is successfully used for various languages [9], its application to German is not sufficiently investigated.

We train an unsupervised model using a vocabulary of unique words that occur more than 5 times in the training data, this gives about 0.5 Million words. We do not include less frequent words in order to avoid noisy words which are harmful to the training process. Nevertheless, the trained model can be used to fragment more unseen words. In addition, the resulting segmentations are adapted to remove noisy fragments, this is found very helpful to improve the final WER.

2.2.3. Graphone-based decomposition

We model the OOV words using *hybrid graphone-augmented word models* which are originally inspired by the work of [10]. The graphone-based component of the model is a sub-lexical model dedicated to deal with OOV words. This approach is strongly related to G2P conversion described in Section 2.1. The set of graphones inferred during G2P training constitutes a graphone-based model that can be easily integrated with the standard word model. This forms a so-called hybrid graphone-augmented word model. Thus, we combine lexical in-vocabulary entries with sub-lexical graphones for OOV words derived from G2P conversion to form a unified set of recognition units. Since the allowed size of graphones is determined via the parameter L (see Section 2.1), we call these *constrained-size graphone-based models*.

Alternatively, we propose a novel version of graphone-based models, where the size constraints are lapsed allowing for *free-size graphones*. Specifically, the full OOV words and

their pronunciations are recovered back from the proposed (constrained-size) graphone sequences. Then, the words are re-split again based on unsupervised Morfessor decomposition to form the graphemic components of the new graphone sequences. While, the phonemic components are obtained after aligning the words to their pronunciations using dynamic programming (DP), and expectation maximization (EM) as proposed in [18]. In other words, we readjust the traditional graphones, which are optimized to give the best *context dependent* pronunciation, by using our unsupervised decomposition, which is expected to produce better fragments suitable for recognition. It is worth noting that practically free-size graphones can be generated by setting L to a large value, but according to [10], this increases the size of the graphone inventory during G2P model training, which causes data sparseness, leading to worse G2P performance.

The main step in building free-size graphones is the letter-phoneme alignment. For that, we need a matrix **A** indexed by all letters and all phonemes, the entries of **A** give the degrees of association between each letter and each phoneme. To find the best alignment, we use two other matrices **B** and **C**. Where, **B** is a matrix of accumulated associations up to some point in the matrix, and **C** holds trace-back pointers indicating from which cell the DP moves. The **B** and **C** matrices are filled left-to-right, top-to-bottom using the following form of recursive maximization equations [18]:

$$B_{i,j} = \max \left\{ \begin{array}{l} B_{i-1,j-1} + A_{l(i),p(j)}, \\ B_{i-1,j}, \\ B_{i,j-1} \end{array} \right\} \begin{array}{l} 1 \leq i \geq L \\ 1 \leq j \geq P \end{array} \quad (4)$$

where L and P are the letter and phoneme lengths of the word respectively. The functions $l(), p()$ provide the letter and phoneme identity at the given index respectively. The entries of **C** are filled according to the chosen maximum. In order to estimate the association matrix **A**, we use the entries of our pronunciation dictionary as training examples along with an EM algorithm that works in the following steps:

- 1. Let $k = 0$, initialize $\mathbf{A}^k$ such that, for every word-pronunciation pair in the dictionary, the entry a^k_{lp} is incremented if the letter l and phoneme p appear in the same pair.
- 2. Use $\mathbf{A}^k$ in DP to align all the word-pronunciation pairs of the dictionary, increment $k = k + 1$
- 3. Compute $\mathbf{A}^k$ such that, for every word-pronunciation pair in the dictionary, the entry a^k_{lp} is incremented if the letter l and phoneme p appear in the same aligned position.
- 4. Go to step 2 until having no change between $\mathbf{A}^k$ and $\mathbf{A}^{k-1}$ (convergence).

2.3. Partial vocabulary decomposition and OOV rate

As previously shown in our earlier work [3], It is useful for sub-lexical LMs to not decompose the top N most frequent decomposable full-words. This prevents the most important words from being confused with other less frequent fragments. Here, we optimize the value of N over the development corpus. In addition, we compute the OOV rate of any test set such that a word is considered an OOV if and only if it is not found in the vocabulary and it is not possible to compose it using vocabulary fragments.

2.4. Word recombination

To allow for easy and deterministic recovery to full-words in the recognition output, we attach a '+' sign to the end of every non-boundary fragment. An example is the word *'absicherung'* which is decomposed into *'ab+ sicher+ ung'*.

3. EXPERIMENTAL SETUP

Our acoustic models are triphone models trained using about 343h of audio material taken from broadcast news (BN), European parliament plenary sessions (EPPS), read articles, dialogs, and some web data. The acoustic models are trained based on maximum likelihood (ML) method.

Our LM training corpora consist of around 306 Million running full-words including data from TAZ newspaper, and web collected German news articles. The vocabularies are selected out of the text corpora by choosing the N top most frequent words, where N ranges from 100k to 300k. The same text corpora are used to estimate back-off N-gram LMs by the SRILM toolkit [19].

Our speech recognizer works in 2 passes. In the first pass, across-word acoustic models are used with no speaker adaptation. The second pass uses the same acoustic models with speaker adaptation based on both Constrained Maximum Likelihood Linear Regression (CMLLR), and Maximum Likelihood Linear Regression (MLLR). In each pass, a 4 or 6-gram LM is used to construct the search space.

To evaluate the recognition performance, we use the Quaero 2009 development and evaluation corpora (dev09: 7.5h; eval09: 3.8h). Each corpus consists of audio material from EPPS sessions and web sources. Additionally, eval09 has some BN data.

4. EXPERIMENTS

4.1. Baseline recognition experiments

In Table 1, we summarize the results of our baseline recognition experiments using traditional LMs based on full-words.

Table 1. *Baseline word error rates [%] using language models based on full-words (voc: vocabulary).*

voc size	Dev09		Eval09	
	OOV [%]	**WER** [%]	**OOV** [%]	**WER** [%]
100k	4.98	33.85	4.79	29.73
200k	3.76	32.67	3.53	28.79
300k	3.26	32.19	2.99	28.36

4.2. Sub-lexical language models based on supervised and unsupervised decomposition

In Table 2, we summarize the results of our recognition experiments using sub-lexical LMs based on supervised word decomposition as shown in Section 2.2.1. The total vocabulary size is fixed to 100k, and the LM order is fixed to 4-gram. We get the best results with supervised decomposition using a vocabulary of *40k full-words + 60k fragments*. We achieve limited WER reductions of [dev09: 0.13% absolute (0.38% relative); eval09: 0.29% absolute (0.98% relative)] compared to the 100k baseline in Table 1. While, no improvement could be achieved over higher vocabulary baseline systems.

Table 2. *Word error rates [%] for sub-lexical LMs based on supervised decomposition (frgs: fragments, wrds: words).*

corpus	#full wrds	# frgs	OOV [%]	WER [%]
Dev09	10k	90k	4.45	34.39
	20k	80k	4.55	34.21
	30k	70k	4.62	33.73
	40k	60k	4.71	**33.72**
	50k	50k	4.77	33.78
	60k	40k	4.77	33.85
Eval09	40k	60k	4.48	**29.44**

In Table 3, we record the recognition results using sub-lexical LMs based on unsupervised decomposition described in Section 2.2.2. We use a vocabulary size of 100k, and a 4-gram LM. We can see that almost all the recognition results are even better than the 300k baseline system. We achieve the best results for unsupervised decomposition using a vocabulary of *5k full-words + 95k fragments*. Significant WER reductions of [dev09: 2.19% absolute (6.47% relative); eval09: 1.28% absolute (4.31% relative)] are achieved compared to the 100k baseline system. On the other hand, compared to the 200k baseline, we get WER reductions of [dev09: 1.01% absolute (3.09% relative); eval09: 0.34% absolute (1.18% relative)]. And compared to the 300k baseline, we get WER improvement of [0.53% absolute (1.64% relative)] for the dev09 corpus, and only slightly worse results for the eval09 corpus.

In Table 4, we record more recognition experiments using larger vocabulary sizes (200k and 300k), and a higher order LM (6-gram). We use the best system configuration as previously optimized in Table 3, fixing the number of full-words to 5k. We can see that using a 6-gram LM does not help. This is caused by the poor LM probability estimates due to sparse data problem. Using a vocabulary of 200k, we get more WER reduction of [dev09: 0.67% absolute (2.17% relative); eval09: 0.21% absolute (0.74% relative)] compared to the 300k baseline. For a 300k vocabulary, we achieve a little more improvement in WER [dev09: 0.76% absolute (2.36% relative); eval09: 0.29% absolute (1.02% relative)] compared to the 300k baseline.

Table 3. *Word error rates [%] for sub-lexical LMs based on unsupervised decomposition.*

corpus	#full wrds	# frgs	OOV [%]	WER [%]
Dev09	0	100k	2.99	32.18
	2k	98k	3.00	31.76
	5k	95k	3.02	**31.66**
	7k	93k	3.04	31.70
	10k	90k	3.07	31.76
	20k	80k	3.19	31.76
	30k	70k	3.32	31.85
Eval09	5k	95k	2.76	**28.45**

Table 4. *More word error rates [%] for sub-lexical LMs based on unsupervised decomposition (#full-words = 5k).*

corpus	voc size	LM	OOV [%]	WER [%]
Dev09	100k	6-gram	3.02	31.62
	200k	4-gram	2.62	31.49
	300k	4-gram	2.40	**31.43**
Eval09	100k	6-gram	2.76	28.48
	200k	4-gram	2.33	28.15
	300k	4-gram	2.13	**28.07**

4.3. Graphone-augmented word models

In Table 5, we record the recognition results using graphone-augmented word models illustrated in Section 2.2.3. The original word model consists of 100k full-words. The LM order is set to 4 or 6-grams, while the graphone size limit L ranges from 2 to 4. We could not set L to higher values as this increases the graphone inventory leading to impractically very large resource requirements. Nevertheless, we optimize L over the dev09 corpus. We see that the 6-gram LM does not improve over the 4-gram LM. Using 4-gram LM with $L = 4$, we achieve WER reductions of [0.29% absolute (0.89% relative)] for dev09 compared to the 200k baseline. For eval09, we can improve WER only over the 100k baseline [0.34% absolute (1.14% relative)].

Table 5. *Word error rates [%] for graphone-augmented word models; graphones are derived from G2P; (L: graphone size limit, gps: graphones); vocabulary size is 100k + #gps.*

corpus	L	# gps	OOV [%]	*4-gram LM* **WER** [%]	*6-gram LM* **WER** [%]
Dev09	2	2k	0.12	34.24	33.00
	3	10k	0.12	32.83	32.72
	4	24k	0.12	**32.38**	32.43
Eval09	4	24k	0.07	**29.39**	29.38

In Table 6, we record the recognition results using graphone-augmented word models. Where, as shown in Section 2.2.3, we use Morfessor decomposition to adjust graphones generated based on $L = 2$, since this gives the least phoneme error rate (PER) for the G2P model. The vocabulary is composed of 100k full-words plus graphones. We can see that the 6-gram LM improves a little over the 4-gram LM. Using the adjusted graphones with a 6-gram LM, by adding 200k graphones to the original vocabulary, we get a little reduction in dev09 WER over the 300k baseline [0.06% absolute (0.19% relative)]. For eval09 WER, we can only improve over the 100k baseline by [0.5% absolute (1.68% relative)]. Furthermore, some experiments are conducted without the phonemic components of graphones, but as results are comparatively worse, we do not include them.

Table 6. *Word error rates [%] for graphone-augmented word models; graphones are adjusted using Morfessor "free-size graphones"; vocabulary size is 100k + #gps.*

corpus	# gps	OOV [%]	*4-gram LM* **WER** [%]	*6-gram LM* **WER** [%]
Dev09	77k	2.80	32.46	32.51
	200k	1.91	32.14	**32.13**
Eval09	77k	2.61	29.49	29.52
	200k	1.71	29.31	**29.23**

4.4. Effect on OOV words

For further analysis of our approaches, we investigate the effect on the recognition of OOV words which are not included in the baseline full-words vocabulary. Therefore, the recognition output is aligned with the reference transcripts and the recognition accuracy is measured only for the OOV region. In Figure 1(a), we show word accuracies using different methods of word decomposition. It can be seen that the largest effect on OOV words occurs with the unsupervised fragments which seem to be very close to the true linguistic morphemes. Thus, by using those fragments we can recognize around 35% of the total OOV words. The differences among other approaches seem not significant. Figure 1(b) provides the recognition accuracies for OOV words measured for larger vocabularies in case of unsupervised fragments. We can see that the same range of accuracy (around 35%) happens for all vocabulary sizes. This indicates the robustness of this approach.

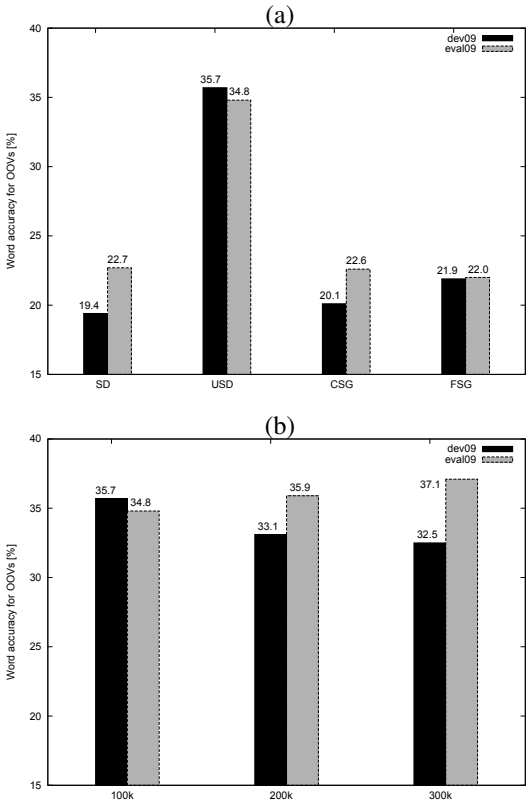

Fig. 1. Recognition accuracies [%] for OOV words; (a) (SD: supervised decomposition with 40k full-words, USD: unsupervised decomposition with 5k full-words, CSG: constrained-size graphones L=4, FSG: free-size graphones 200k) using vocabulary size = 100k and 4-gram LMs; (b) unsupervised decomposition using 100k, 200k and 300k vocabularies and 4-gram LMs.

5. CONCLUSIONS

We have investigated the use of sub-lexical language models for German ASR. Three approaches for vocabulary decomposition are compared, namely supervised and unsupervised word decomposition, in addition to the graphone-based decomposition. The best approach is to use a vocabulary of fragments generated by unsupervised methods, along with some fraction of full-words (around 5k). In our experiments, we have shown that using a vocabulary size of only 100k *(5k full-words + 95k fragments)*, we could significantly improve WER over a 100k system of full-words by [dev09: 2.19% absolute (6.47% relative); eval09: 1.28% absolute (4.31% relative)]. Increasing the overall vocabulary size decreases the WER correspondingly. The improvement gets less for higher vocabulary sizes, the reason is the higher degree of acoustic confusability introduced by short fragments. The graphone-augmented word models perform better than the supervised decomposition approach, but worse than the unsupervised decomposition approach. Moreover, a novel method is introduced which uses word decomposition to allow for free-size graphones. We believe that if a standard decomposition is available, we can further reduce the WER. Ongoing work includes the use of free-size graphones instead of the normal fragments in the original recognition vocabulary. One choice is to consider a 100k vocabulary of 5k full-words + 95k morfessor graphones, where graphones replace the traditional graphemic fragments.

6. ACKNOWLEDGEMENTS

This work was partly funded by the European Community's Seventh Framework Programme under the project SCALE (FP7-213850), and was partly realized under the Quaero Programme, funded by OSEO, French State agency for innovation.

7. REFERENCES

[1] M. Adda-Decker and G. Adda, "Morphological decomposition for ASR in German," in *Workshop on Phonetics and Phonology in ASR*, Saarbrücken, Germany, Mar. 2000, pp. 129 – 143.

[2] A. Berton, P. Fetter, and P. Regal-Brietzmann, "Compound words in large-vocabulary German speech recognition systems," in *Proc. Int. Conf. on Spoken Language Processing*, Philadelphia, PA, USA, Oct. 1996, vol. 2, pp. 1165 – 1168.

[3] A. El-Desoky, C. Gollan, D. Rybach, R. Schlüter, and H. Ney, "Investigating the use of morphological decomposition and diacritization for improving Arabic LVCSR," in *Interspeech*, Brighton, UK, Sept. 2009, pp. 2679 – 2682.

[4] L. Lamel, A. Messaoudi, and J.L Gauvain, "Investigating morphological decomposition for transcription of Arabic broadcast news and broadcast conversation data," in *Interspeech*, Brisbane, Australia, Sept. 2008, vol. 1, pp. 1429 – 1432.

[5] J. Kneissler and D. Klakow, "Speech recognition for huge vocabularies by using optimized sub-word units," in *Proc. European Conf. on Speech Communication and Technology*, Aalborg, Denmark, Sept. 2001, vol. 1, pp. 69 – 72.

[6] M. Adda-Decker, "A corpus-based decompounding algorithm for German lexical modeling in LVCSR," in *Proc. European Conf. on Speech Communication and Technology*, Geneva, Switzerland, Sept. 2003, pp. 257 – 260.

[7] R. Ordelman, A. V. Hassen, and F. D. Jong, "Compound decomposition in Dutch large vocabulary speech recognition," in *Proc. European Conf. on Speech Communication and Technology*, Geneva, Switzerland, Sept. 2003, pp. 225 – 228.

[8] M. Larson, D. Willett, J. Köhler, and R. Rigoll, "Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches," in *Proc. Int. Conf. on Spoken Language Processing*, Beijing, China, Oct. 2000.

[9] M. Creutz, T. Hirsimki, M. Kurimo, A. Puurula, J. Pylkknen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraclar, and A. Stolcke, "Morph-based speech recognition and modeling of out-of-vocabulary words across languages," *ACM Transactions on Speech and Language Processing*, vol. 5, no. 1, Dec. 2007.

[10] M. Bisani and H. Ney, "Open vocabulary speech recognition with flat hybrid models," in *Interspeech*, Lisbon, Portugal, Sept. 2005, pp. 725 – 728.

[11] I. Bazzi and J. R. Glass, "Modeling out-of-vocabulary words for robust speech recognition," in *Proc. Int. Conf. on Spoken Language Processing*, Beijing, China, Oct. 2000.

[12] D. Klakow, G. Rose, and X. Aubert, "OOV-detection in large vocabulary system using automatically defined word-fragments as fillers," in *Proc. European Conf. on Speech Communication and Technology*, Budapest, Hungary, Sept. 1999, vol. 1, pp. 49 – 52.

[13] L. Galescu, "Recognition of out-of-vocabulary words with sub-lexical language models," in *Proc. European Conf. on Speech Communication and Technology*, Geneva, Switzerland, Sept. 2003, pp. 249 – 252.

[14] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434 – 451, May 2008.

[15] P. Koehn and K. Knight, "Empirical methods for compound splitting," in *Proc. the Conference of the European Chapter of the ACL*, Budapest, Hungary, Apr. 2003, pp. 347 – 354.

[16] M. Creutz and K. Lagus, "Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0," Tech. Rep., Computer and Information Science Helsinki University of Technology, Finland, Mar. 2005.

[17] M. Creutz, *Induction of the morphology of natural language: Unsupervised morpheme segmentation with application to automatic speech recognition*, Ph.D. thesis, Helsinki University of Technology, Finland, 2006.

[18] R. I. Damper, Y. Marchand, J. D. Marsters, and A. Bazin, "Aligning letters and phonemes for speech synthesis," in *5th ISCA Speech Synthesis Workshop*, Pittsburg, PA, USA, June 2004, pp. 209 – 214.

[19] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. Int. Conf. on Spoken Language Processing*, Denver, Colorado, USA, Sept. 2002, vol. 2, pp. 901 – 904.